DOI:10.3969/j.issn.1005-202X.2022.01.001

医学放射物理

深度强化学习在直肠癌IMRT自动计划的应用

王翰林,刘嘉城,王清莹,岳海振,杜乙,张艺宝,王若曦,吴昊
北京大学肿瘤医院暨北京市肿瘤防治研究所放疗科/恶性肿瘤发病机制及转化研究教育部重点实验室,北京 100142

【摘要】目的:对于调强放疗(IMRT)计划,优化过程较为耗时,且计划的质量取决于计划人员的经验和时间,本文探讨并实现一种无监督IMRT自动优化的方案,使其能够模拟人工操作方式进行治疗计划优化。**方法:**本研究基于深度强化学习框架,提出一种优化调整决策网络(OAPN)自动化计划优化的方法。利用Varian Eclipse 15.6 TPS的脚本应用程序接口(ESAPI)实现OAPN与TPS之间的交互,以剂量体积直方图作为信息输入,通过强化学习的训练方式,OAPN学习TPS中目标参数的调整策略,从而逐步改善并获得较高质量的计划。实验从临床数据库中选取18例既往已完成治疗的直肠癌病例,其中5例计划案例用于OAPN网络训练,其余13例计划案例用于评估训练后OAPN的可行性与有效性,引入第三方计划评分工具来衡量计划质量。**结果:**用于测试的13例直肠癌计划,使用统一的初始优化目标参数(OOPs)所获得的平均得分为(45.53±4.58)分(计划得分上限值为110),经过OAPN对OOPs调整后计划所获得的平均得分为(88.67±6.74)分。**结论:**OAPN借助ESAPI实现与TPS之间数据交互,通过深度强化学习的方式形成行为价值策略,经过训练后的OAPN可以对目标参数进行高效率的调整,同时获得较高质量计划。

【关键词】直肠癌;自动优化;深度强化学习;脚本应用程序接口;优化调整决策网络
【中图分类号】R318;R811.1 **【文献标志码】**A **【文章编号】**1005-202X(2022)01-0001-08

Application of deep reinforcement learning in automatic IMRT planning for rectal cancer

WANG Hanlin, LIU Jiacheng, WANG Qingying, YUE Haizhen, DU Yi, ZHANG Yibao, WANG Ruoxi, WU Hao
Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education)/Department of Radiotherapy, Peking University Cancer Hospital & Institute, Beijing 100142, China

Abstract: Objective The optimization of intensity-modulated radiotherapy planning is often time-consuming, and the plan quality depends on the experience of the planner and the available planning time. An unsupervised automatic intensity-modulated radiotherapy optimization procedure is discussed and implemented to simulate the human operation during the whole optimization process. **Methods** Based on the framework of deep reinforcement learning (DRL), an optimization adjustment policy network (OAPN) was proposed to automate the process of treatment planning optimization. The scripting application programming interface (ESAPI) of Varian Eclipse 15.6 TPS was used to realize the interaction between OAPN and TPS. Taking dose-volume histogram as the information input, OAPN learned the adjustment strategy of objective parameters in TPS by the training mode of reinforcement learning, so as to gradually improve and obtain high-quality plans. A total of 18 cases of rectum cancer which had completed treatment were selected from the clinical database. Five of the cases were used for OAPN training, and the remaining 13 for evaluating the feasibility and effectiveness of OAPN after training. Finally, a third-party scoring tool was used to evaluate plan quality. **Results** The average score of 13 tested plans using uniform initial optimization objective parameters (OOPs) was 45.53±4.58 (the upper limit value was 110). After adjusting OOPs by OAPN, the average plan score was 88.67±6.74. **Conclusion** OAPN can realize the data interaction with TPS through ESAPI, and form an action-value strategy through DRL. After training, OAPN can efficiently adjust OOPs and obtain a high-quality plan.

Keywords: rectum cancer; automatic optimization; deep reinforcement learning; Eclipse scripting application programming interface; optimization adjustment policy network

【收稿日期】2021-07-10
【基金项目】国家重点研发计划(2019YFF01014405);北京市医管局培育计划(PX2019042);北京市自然科学基金(1202009);国家自然科学基金(12005007);中央高校基本科研业务费/北京大学临床医学+X青年专项(PKU2020LCXQ019)
【作者简介】王翰林,硕士,研究方向:放射治疗物理学,E-mail: Wanghanlins@163.com
【通信作者】吴昊,高级工程师,研究方向:医学物理,E-mail: hao.wu@bjcancer.org

前 言

目前，基于逆向优化方法的调强放疗（Intensity-Modulated Radiation Therapy, IMRT）已成为放射治疗的主要技术手段[1]。针对各种临床实际情况，逆向优化参数通常包含多项约束条件，例如靶区和危及器官的剂量学指标和权重等。物理师在计划设计过程中通常会根据剂量体积直方图（Dose Volume Histogram, DVH）和剂量分布信息，通过试错的方式在优化过程中不断调整约束参数，从而不断改善计划质量，在给予肿瘤致死剂量的同时尽可能减少周围正常组织的损伤[2]。治疗计划质量优化所耗费的时间会因物理师个人经验和治疗计划的复杂程度出现较大的差异。为了提高计划质量和一致性、缩短计划设计所耗费的人工时间，一些研究者提出了基于不同理论的自动计划方法[3]：基于经验的计划设计（Knowledge-Based Planning, KBP）[4-5]；基于协议的自动迭代优化（Protocol Based Automatic Iterative Optimization, PB-AIO）[6-7]；多标准优化（Multi-Criteria Optimization, MCO）[8-9]等。自动计划的相关研究促使近年来放疗计划设计过程朝着自动化与标准化方向发展，部分技术还陆续实现了商品化。

强化学习近年来在解决一些困难的顺序决策问题中取得了巨大的成功。其中最具有亮点的例子为：在众多Atari游戏项目中获取了超过人类专家的得分[10]；以及Alpha Go在围棋上的突破[11]。放疗计划设计一定程度上也是一个顺序决策的过程，但将强化学习应用于放疗计划设计的研究鲜有报道[12]。不同于KBP、PB-AIO等方法，强化学习不需要利用既往经验设计模型，而是将治疗计划系统（Treatment Planning System, TPS）反馈的信息（包括剂量分布、DVH等）直接传递给智能体（Agent），Agent通过预先定义的评分标准对计划结果进行评分。在多次循环过程中，Agent学习调整约束限值与计划评分改变之间的潜在联系，并在多个计划优化过程中将该联系泛化，最后将学习到的潜在联系应用到新的计划优化过程中。在学习过程中，Agent设计是通用的，不需要人为提取计划设计的先验知识，能够大大提高自动计划应用的效率和场景。

本研究基于Deep Q Learning[10]的方法提出了一种优化调整决策网络（Optimization Adjustment Policy Network, OAPN），利用Eclipse TPS 15.6（Varian Medical Systems, Palo Alto, CA, US）的脚本应用程序接口（Eclipse Scripting Application Programming Interface, ESAPI）进行临床计划的优化。实验结果表明OAPN经过训练学习到的调整策略，展现出提升计划质量的潜力。

1 方 法

1.1 病例资料

随机选择北京大学肿瘤医院放疗科2019年9月~2020年8月已经完成放疗的18例直肠癌术前病例。患者年龄40~70岁，中位年龄为61岁，定位CT层厚为5 mm。本次实验的所有病例均采用相同的临床处方剂量标准，计划肿瘤靶区（PGTV）50.6 Gy/25次，计划靶区（PTV）41.8 Gy/25次（PGTV和PTV为肿瘤靶区和临床靶区外扩5 mm得到），要求靶区95%以上的体积接受的剂量应达到处方剂量。

1.2 治疗计划优化算法

本研究使用Eclipse TPS 15.6的光子优化（Photon Optimizer, PO）[13]算法，该算法替代了之前用于固定野IMRT和容积调强放疗（Volumetric Modulated Arc Therapy, VMAT）的Dose Volume Optimizer算法与Progressive Resolution Optimizer算法。目前，Eclipse的优化过程是基于多种DVH约束条件下进行的。优化过程中可设置的约束条件类型较多，如体积、剂量、优先权、正常组织目标（Normal Tissue Objective, NTO）、MU目标等，其中最为常用的约束项为体积剂量约束。一个体积剂量约束对应一个优化目标参数（Optimization Objective Parameters, OOPs）集合，OOPs主要由3部分组成：优化目标权重、表示DVH图中剂量-体积的二维坐标、在DVH曲线上约束条件的方向（大于/小于）。PO算法根据给定OOPs集合，构建对应的优化目标函数，通过调整计划参数最小化目标函数，得到最终解。本研究为了简化强化学习问题中的约束类型，仅使用了基于DVH曲线中体积剂量的约束项。

1.3 OAPN

针对上述优化算法，本文提出了一种OAPN来实现计划优化过程的自动化。初始参数设置参考了临床计划过程中相对宽松的OAR体积剂量约束条件，以扩展待搜索的整体参数空间。与物理师在治疗计划设计中的人为操作类似，OAPN通过ESAPI获取优化状态并进行判断后，输出对OOPs的调整策略（包括调整的方向和幅度），并利用ESAPI将更新的OOPs输入TPS继续进行优化。如图1所示，此过程持续进行直到OOPs调整次数达到最大限值为止。

上述过程可以归类于强化学习中主要解决的策略问题：在一系列环境状态下，Agent通过选择行为，使得最终的评估结果收益最大化或最优化。其中，Qlearning[14]作为一种基于值函数的强化学习算法，其训练过程依靠Bellman方程[15]定义最优行为价值函数，并按照最优策略执行以获得最大收益。此行

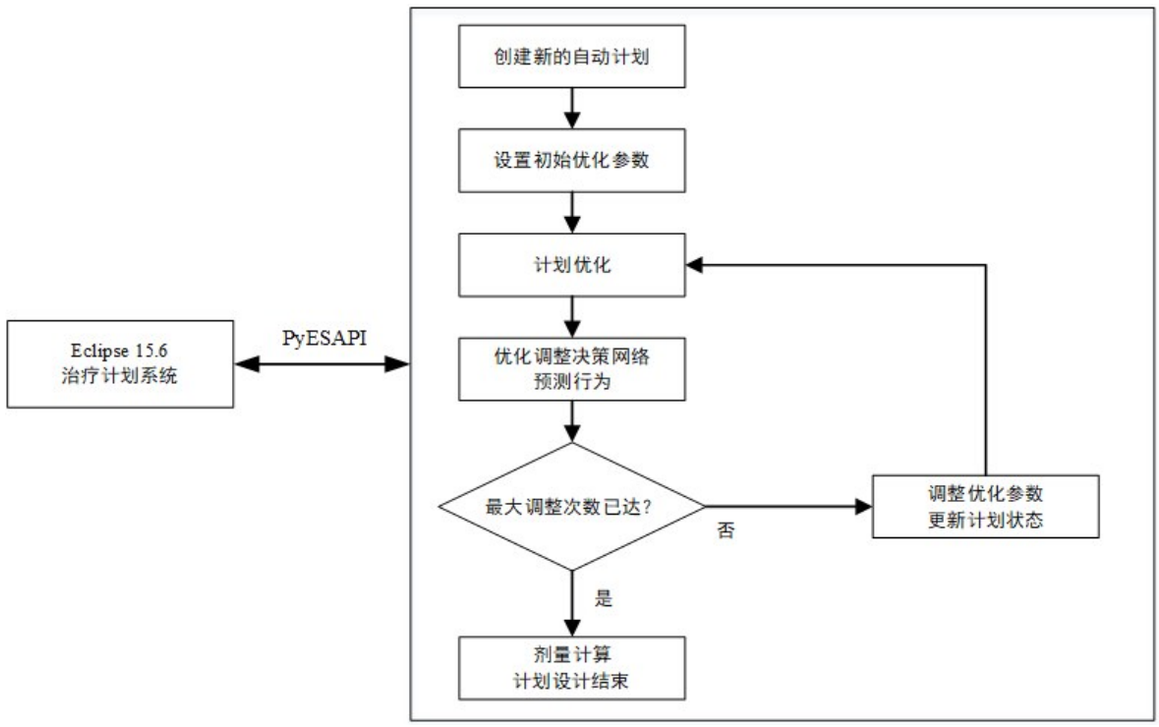


图1 基于优化调整决策网络(OAPN)的自动IMRT计划工作流程图

Figure 1 Workflow of the automatic IMRT planning based on optimization adjustment policy network (OAPN)

为价值函数一般形式定义如下：

$$Q^*(s,a)=r+\gamma\max_a Q^*(s',a')\big|s^n=s,a^n=a,s^{n+1}=s',a^{n+1}=a' \quad (1)$$

如果将公式(1)中的$Q^*(s',a')$也按上式展开，以此类推，最优行为-价值函数表达为：

$$Q^*(s,a)=\max_\pi\left[r^n+\gamma r^{n+1}+\gamma r^{n+2}+\cdots\big|s^n=s,a^n=a,\pi\right] \quad (2)$$

其中，状态s表示Agent在该次优化中观测到的环境，行为a表示对OOPs的调整。s^n,a^n分别代表OOPs在第n次调整时的状态与行为。r^n表示OOPs在第n次调整后获得的奖励(由预先定义的奖励函数给出)，如果在状态s^n下采取行为a^n可以获得更好的计划则奖励大于0，否则奖励小于0。$\gamma\in[0,1]$为衰减因子，用来调节行为-价值函数对当前奖励和未来奖励的重视程度。$\pi=P(a|s)$代表OOPs的调整策略(基于观测状态s采取行为a)。优化过程中OOPs的自动调整是以建立$Q^*(s,a)$函数为目标，一旦$Q^*(s,a)$函数确定，则选择在观测状态s下取得最大Q值的行为作为行为策略，即$a=\arg\max_a Q^*(s,a')$。

为了确定$Q^*(s,a)$函数的一般形式，本研究搭建深度学习网络OAPN，如图2所示，OAPN包含一个输入层(大小为600×3)，7个全连接层，6个ReLU层和一个输出层(大小为6×1)，Loss函数为均方误差，优化器为Adam(Learning Rate=0.000 1)。利用靶区和危及器官的DVH参数作为OAPN的输入(如PTV、股骨头和膀胱等)。在计划优化中，OAPN会对其OOPs进行调整。对于每个需要被调整的OOPs，共有2种OOPs调整行为，即增加10%和降低10%。则OAPN共有6个输出结果，分别是不同行为所对应的Q值。

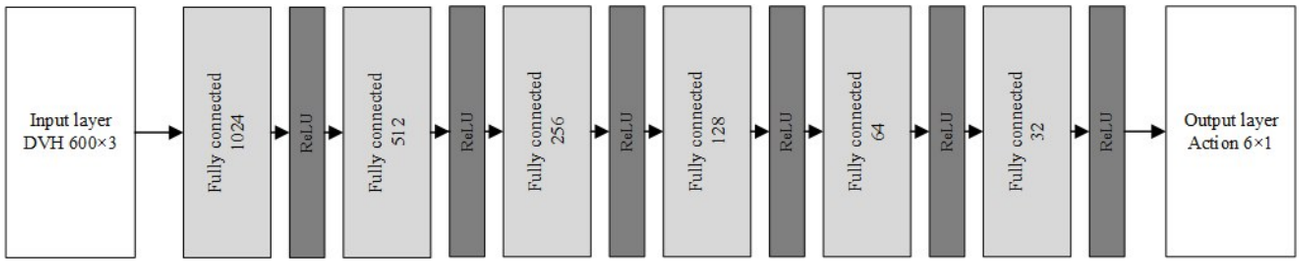


图2 OAPN结构图

Figure 2 OAPN structure

1.4 奖励函数

本研究通过定义得分函数φ来量化计划质量，因此，OAPN可以根据奖励函数定量的评估OOPs调整后计划质量的改变。使用Sun Nuclear公司的Plan IQ软件为直肠癌IMRT计划提供一系列评估指标。对于每个指标，得分被定义为0~10之间的分段线性函数。与Shen等[16]研究相似，本文采用平滑的Sigmoid-shape函数代替分段线性函数，以改善Plan IQ阶跃式惩罚函数的局限性。根据靶区剂量覆盖和OAR剂量控制的不同临床要求，得分函数φ定义了11个等权重的评分指标，包括PTV D_1，膀胱V_{15}、V_{20}、V_{25}、V_{30}、V_{35}，股骨头V_5、V_{10}、V_{15}、V_{20}、V_{25}。最终的计划得分将被计算为所有指标评分的总和，得分越高表示计划质量越好，一个计划可达到的最高分数是110分。则奖励函数的定义如下：

$$r = \varphi\left(s^{n+1}\right) - \varphi\left(s^{n}\right) \quad (3)$$

奖励函数明确地衡量了在OOPs调整前后状态s^n和s^{n+1}计划质量的差异。如果$r > 0$，表示计划质量得到提高，反之计划质量降低。

1.5 经验回放和固定Q目标

通过OAPN网络训练获取$Q(s,a;W)$的一般形式，OAPN网络参数W通过训练不断更新。为了提高训练的稳定性和收敛性，本文将experiencereplay和fixed Q target的策略[17]加入网络参数训练，其损失函数定义如下：

$$H(W)=\frac{1}{N}\sum_{i=1}^{N}\left[r+\gamma\max_{a}Q\left(s',a';\hat{W}\right)-Q(s,a;W)\right]^2 \quad (4)$$

其中，N表示训练集的大小。研究的目标是通过OAPN确定参数W，使损失函数$H(W)$最小化。网络参数通过experiencereplay策略进行更新，每次随机选取15个训练样本，并使用Adam优化器最小化损失函数$H(W)$。对于Q目标网络参数$\hat{W}$，在训练过程中保持相对固定，并每隔20次训练周期更新为最新网络参数W。

1.6 网络训练

训练过程如算法1所示，使用5个直肠癌IMRT计划在100个episode（指Agent从开始进行行为调整到调整终止的一次完整过程）下进行，对于每个episode，训练计划会初始化统一的OOPs，并进行一系列OOPs调整步骤，每一步都包含一个计划优化的过程。如果OOPs调整次数达到最大限值30次，OOPs会终止调整，然后转移到下一个训练病例。在每一步中，本文引入ϵ - greedy算法[18]去选择一个行动来调整OOPs。概率ϵ在训练过程中以0.95/episode的衰减率衰减，以调整训练的OAPN对探索与利用两种行为的配比。在行为被选择后，OOPs会被调整并执行一次计划优化，利用OOPs调整前后计划的DVH曲线计算奖励函数r。OOPs调整前后的状态s^n和s^{n+1}，选择的行为a^n和奖励r^n共同构成一个训练样本。重复以上过程生成大量的训练样本。利用experiencereplay和fixed Q target策略进行OAPN训练，从而可以克服顺序生成的训练样本之间的强相关性。最终，参数W和$\hat{W}$随着训练逐渐收敛。

算法1 标准DRL算法训练OAPN

输入：$N_{episode} = 100$
$N_{patient} = 5$
$N_{train} = 30$
$N_{batch} = 15$
1. 搭建OAPN网络框架（图2），初始化网络参数
for episode = 1, 2, ⋯, $N_{episode}$ ***do***
  for p = 1, 2, ⋯, $N_{patient}$ ***do***
    2. 初始化TPPs
      将初始化TPPs带入TPS中优化，获得状态s^0
    for n = 1, 2, ⋯, N_{train} ***do***
      3. 利用ϵ − greedy选择行为a^n:
        Case1:ϵ的概率，随机选择a^n
        Case2:1-ϵ的概率，$a^n = \arg\max_a Q^*$
      4. 利用a^n更新TPPs
      5. 将更新TPPs带入TPS中优化，获得状态s^{n+1}
      6. 计算奖励$r^n = \varphi\left(s^{n+1}\right) - \varphi\left(s^{n}\right)$（公式3）
      7. 向训练数据池中存储训练样本$\{s^n, a^n, r^n, s^{n+1}\}$
      8. 利用experiencereplay训练网络参数
        从训练数据池中随机选取N_{batch}训练样本
    end for
  end for
end for
输出：网络参数

在本研究中，使用Python语言和Keras框架实现OAPN网络搭建和模型训练，并基于Eclipse TPS 15.6自带的ESAPI脚本接口完成OAPN的训练和与TPS的交互。整体过程在一台桌面工作站实现（双Intel Xeon 3.1 GHz CPU处理器，Nvidia Quadro P5000 GPU卡）。

1.7 模型评估

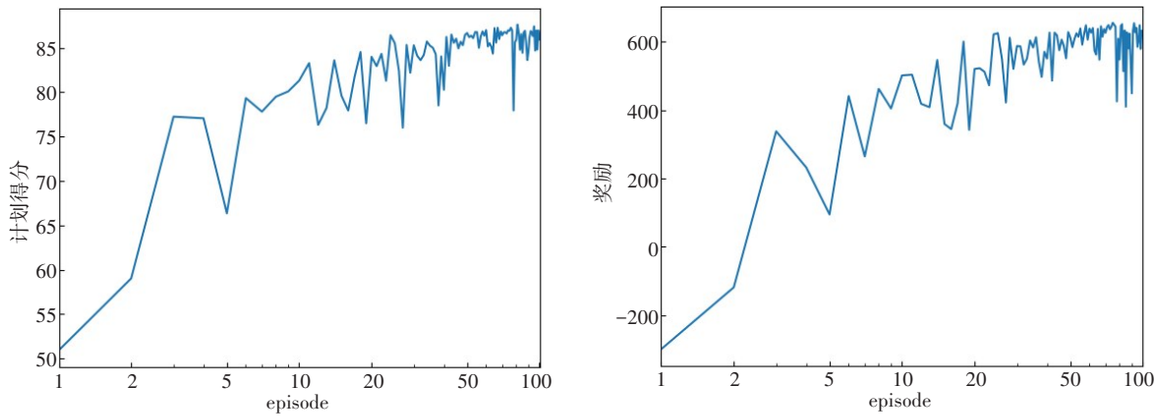
收集临床数据库中18例直肠癌患者，其中5例计划用于OAPN训练，其余13例计划用于评估训练过的OAPN模型性能。对于每个计划，所有OOPs的初始化设置为统一值。按照图1的流程进行迭代调整，如果OAPN调整次数达到最大值40次，则迭代终止。在所有按照OAPN指导的OOPs调整依次生成的中间计划中，选择得分最高的中间计划作为最终计划。将最终计划的DVH曲线和计划得分φ与使用初始OOPs生成的初始计划进行比较，以评估计划的质量的改进。

2 结　果

2.1 训练集

在本实验中，5例直肠癌IMRT计划作为OAPN训练集进行网络训练，训练总体用时大约35 h。图3、图4为1例训练案例，展示计划得分、奖励和Q值在网络训练中不断变化的过程。

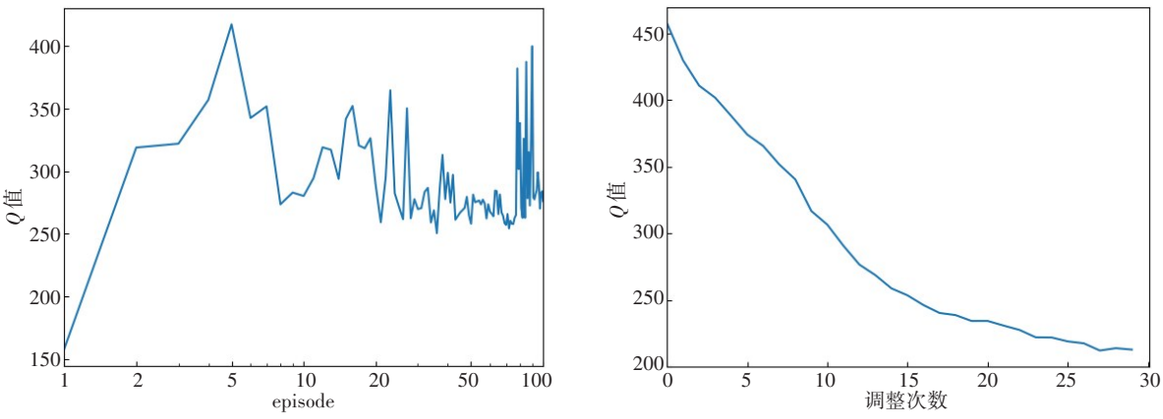
计划得分和奖励随着训练进行的变化趋势如图3所示。其中，计划得分和奖励反映使用VTPN自动调整OOPs所获得的计划质量的改善，计划得分和奖励总体呈上升趋势，其中计划得分约提升60%，与最初的计划相比，最终的计划质量有了很大的提高。图4a记录下Q值在100个episode中的变化过程，Q值总体呈上升趋势，表示OAPN的表现性能逐步提升。Q值在每个episode下的30次调整过程中平均变化如图4b所示，经过30次调整后，Q值约降低为初始的一半。Q值在30次调整过程中的持续下降，表示随着OOPs的不断调整，计划质量在提升，但计划难度正在增加，未来价值在变小。上述结果表明OAPN在学习如何调整OOPs来提高计划质量。



a：计划得分　　b：奖励

图3 训练案例在训练过程中计划得分和奖励的变化图

Figure 3 Trends of plan score and reward during training for one training case



a：episode　　b：调整次数

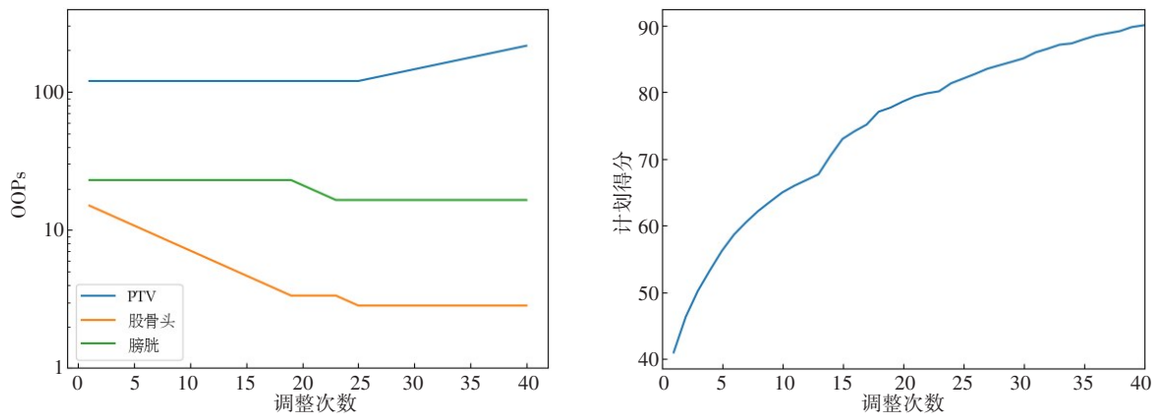
图4 训练案例在训练过程中Q值变化图

Figure 4 Trends of Q value during training for one training case

2.2 测试集

本研究将没有参与网络训练的13例直肠癌计划案例用于评估训练后OAPN的可行性与有效性，并选取1个代表性案例进行结果展示。

如图5所示，OAPN首先决定下调股骨头OOPs中的平均剂量，从而控制股骨头的剂量，随后OAPN对膀胱的OOPs进行调整，成功的减少了膀胱的剂量，最后，OAPN决定增加PTV靶区的权重，以控制PTV内的高剂量区。在这个过程中，该计划的Plan IQ得分从初始的40分增长至90分以上。图6展示出DVH曲线在OOPs调整过程中的变化情况，股骨头和膀胱的DVH曲线随着调整次数的增加有着明显下降的趋势，而PTV利用初始OOPs获得的DVH曲线与在第15次、25次和40次进行OOPs调整后获得的DVH曲线没有显著变化。以上的结果表明OAPN学习到的调整OOPs策略与临床中人工计划优化思路相类似，计划质量有明显提升。



a：OOPs　　b：计划得分

图5 代表性案例在计划优化过程中优化目标参数（OOPs）的调整过程

Figure 5 Adjustments of optimization objective parameters (OOPs) during planning optimization for one representative case

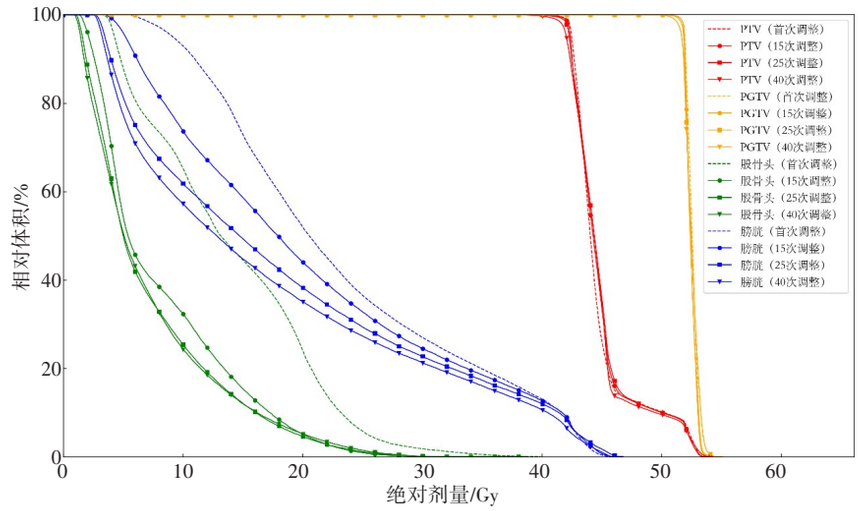


图6 代表性案例在OOPs调整过程中DVH曲线的变化情况

Figure 6 Trends of DVH curves during OOPs adjustment for one representative case

OAPN在所有测试计划案例中的表现如表1所示。初始OOPs在经过OAPN调整后，计划得分从（45.53±4.58）分提升至（88.67±6.74）分，股骨头和膀胱的平均剂量分别下降了53.1%和29.2%，PTV和PGTV平均$D_{95\%}$指标也满足大于处方剂量的要求（PTV $D_{95\%}$>41.8 Gy；PGTV $D_{95\%}$>50.6 Gy）。利用OAPN在所有测试案例进行计划设计的平均耗时约为4 min，其中大部分时间用于每次调用TPS优化引擎进行计划优化的过程。

3 讨 论

本研究在Eclipse计划系统平台上开发了一种直

表1 初始OOPs和第20次、第40次OOPs调整后所获得的计划参数对比($\bar{x} \pm s$)

Table 1 Comparison of planning parameters of the original OOPs and those after the 20th and 40th OOPs adjustments (*Mean±SD*)

调整次数	计划得分/分	股骨头 D_{mean}/Gy	膀胱 D_{mean}/Gy	PTV $D_{95\%}$/Gy	PGTV $D_{95\%}$/Gy	PTV $D_{1\%}$/Gy
初始OOPs	45.53±4.58	13.91±1.35	24.36±1.92	42.54±0.18	51.97±0.16	53.17±0.13
20次调整后	79.95±4.27	8.38±1.45	17.63±2.17	42.41±0.27	51.85±0.12	53.34±0.31
40次调整后	88.67±6.74	6.51±1.66	17.23±1.98	42.20±0.27	51.73±0.13	53.56±0.32

肠癌IMRT自动优化方法。该方法基于深度强化学习框架,搭建了OAPN网络,实现了网络训练和自动优化的任务。与大多数基于KBP和PB-AIO方法的自动优化不同,本工作采用离线差分学习的策略,通过定义奖励函数,让Agent在计划系统的环境中学习OOPs调整的行为价值函数,从而实现IMRT治疗计划优化的自动化。同时使用一个深层卷积网络将对应DVH参数转化为强化学习的输入特征,避免了DVH中的人工特征提取工作。该方法的训练是通过Agent在计划系统中的大量迭代进行的,不需要积累大量的以往优质计划,因此该方法在多中心的推广有较大意义。此外,经过网络训练得到的OAPN模型,可以同时在后台执行多个IMRT自动计划,可缩减计划设计中的人力和人工耗时。

基于本研究的网络框架,在TPS工作站上训练OAPN大约需要35 h。其中,主要的计算工作花费在OAPN网络训练中重复执行IMRT优化过程,由训练所用病例数、OOPs调整步骤和训练episode数所决定。在本研究中,为快速测试方法的可行性,训练病例的数量限制为5个,而最大OOPs调整步骤和训练episode分别设置为30和100。为了提高OAPN的通用性和鲁棒性,需要使用更大的训练样本进行验证。在未来的工作中,我们希望利用该方法处理更加复杂的临床问题,例如:更多样化的OOPs调整选择或更复杂放疗计划类型等。

在本研究中,利用了Plan IQ为直肠癌IMRT计划提供一系列评估指标,虽然OAPN已经展现出具有提升计划质量的潜能,但仅仅基于Plan IQ计划得分的奖励函数不能全面反映计划质量评价的临床标准。细化和完善现有的计划质量评价标准将是下一步工作的重点,例如增加对靶区剂量覆盖和OAR的剂量控制方面的评价标准可以更好的适应临床的需要。在未来工作中,我们将尝试量化临床物理师的判断以更好的反映奖励函数,并将临床中更多的判断指标纳入OAPN的训练中。

此外,虽然DVH通常是临床上最为关注的计划质量评估方法,但由于其无法提供剂量空间信息,OAPN仅仅根据DVH曲线来进行网络训练和计划质量评判有待进一步完善。如图6中,PTV利用初始OOPs获得的DVH曲线与在第15次、25次和40次进行OOPs调整后获得的DVH曲线并没有明显的变化,$D_{95\%}$和$D_{1\%}$均可以满足我们的要求,但PTV的空间剂量信息却无法从DVH曲线中得以评估,例如靶区冷热点问题、靶区剂量适形性,靶区剂量均匀性等等。在未来的工作中,我们考虑将三维剂量图像的完整信息也应用到此方法中,以更加贴合临床的需求。

在本研究中,ESAPI起到了至关重要的作用,其完成OAPN与TPS之间数据交互的同时,也实现了IMRT计划设计的自动化,如果其他TPS存在类似的可编程接口,该网络可以很容易地集成到任何TPS中进行应用。其应用灵活,并可以根据需要拓展至其他病种(如宫颈癌、鼻咽癌等)。因此,它可以在未来工作中作为一个有效的工具来减少不同的物理师,甚至不同的治疗系统之间计划质量一致性的差异。

4 结 论

借助ESAPI脚本接口,本研究在Eclipse平台下开发了一种深度强化学习网络,选取直肠癌放疗计划验证了所提出想法的可行性和潜力,并实现了直肠癌的IMRT治疗计划设计过程中优化的自动化。开发的OAPN能够模拟临床优化调整的行为,学习如何调整优化引擎中的OOPs生成令人满意的计划。在本项工作中,OAPN通过训练自主学习计划优化中的行为-价值策略,初步显示了深度强化学习技术应用于商业治疗计划系统的潜力,可以完成学习和掌握临床中放疗治疗计划优化任务。

【参考文献】

[1] EZZELL G A, GALVIN J M, LOW D, et al. Guidance document on delivery, treatment planning, and clinical implementation of IMRT: report of the IMRT subcommittee of the AAPM radiation therapy committee[J]. Med Phys, 2003, 30(8): 2089-2115.

[2] NELMS B E, ROBINSON G, MARKHAM J, et al. Variation in external beam treatment plan quality: an inter-institutional study of planners and planning systems[J]. Pract Radiat Oncol, 2012, 2(4): 296-305.

[3] HUSSEIN M, HEIJMEN B J, VERELLEN D, et al. Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations[J]. Br J Radiol, 2018, 91(1092): 20180270.

[4] SHIRAISHI S, MOORE K L. Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy[J]. Med Phys, 2016, 43(1): 378-387.

[5] GOOD D, LO J, LEE W R, et al. A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: an example application to prostate cancer planning[J]. Int J Radiat Oncol Biol Phys, 2013, 87(1): 176-181.
[6] SONG Y, WANG Q, JIANG X, et al. Fully automatic volumetric modulated arc therapy plan generation for rectal cancer[J]. Radiother Oncol, 2016, 119(3): 531-536.
[7] ZHANG X, LI X, QUAN E M, et al. A methodology for automatic intensity-modulated radiation treatment planning for lung cancer[J]. Phys Med Biol, 2011, 56(13): 3873-3893.
[8] BREEDVELD S, STORCHI P R, KEIJZER M, et al. A novel approach to multi-criteria inverse planning for IMRT[J]. Phys Med Biol, 2007, 52(20): 6339-6353.
[9] MONZ M, KÜFER K H, BORTFELD T R, et al. Pareto navigation-algorithmic foundation of interactive multi-criteria IMRT planning[J]. Phys Med Biol, 2008, 53(4): 985-998.
[10] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
[11] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529 (7587): 484-489.
[12] SHEN C, GONZALEZ Y, KLAGES P, et al. Intelligent inverse treatment planning *via* deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer[J]. Phys Med Biol, 2019, 64(11): 115013.
[13] LIU H, SINTAY B, PEARMAN K, et al. Comparison of the progressive resolution optimizer and photon optimizer in VMAT optimization for stereotactic treatments[J]. J Appl Clin Med Phys, 2018, 19(4): 155-162.
[14] CHRISTOPHER J. Q-learning[J]. Mach Learn, 1992, 8(3): 279-292.
[15] BELLMAN R E. A markov decision process[J]. J Math Fluid Mech, 1957, 6: 679-684.
[16] SHEN C, NGUYEN D, CHEN L, et al. Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning[J]. Med Phys, 2020, 47(6): 2329-2336.
[17] FANG M, LI Y, COHN T. Learning how to active learn: a deep reinforcement learning approach [C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 595-605.
[18] SUTTON R S, BARTO A G. Reinforcement Learning: An Introduction [J]. IEEE Trans Neural Netw, 1998, 9(5): 1054.

（编辑：薛泽玲）